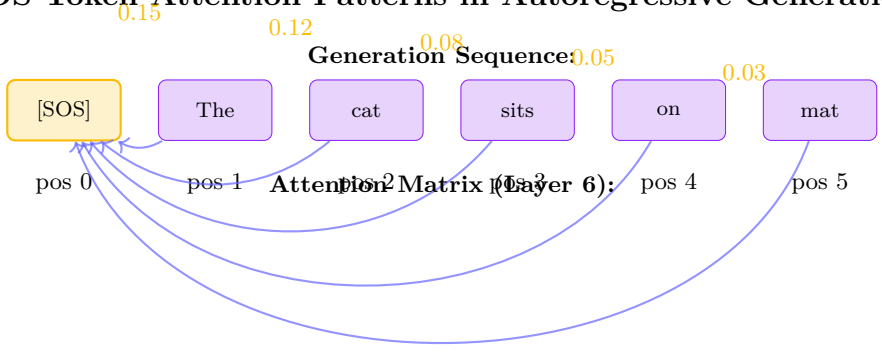
SOS Token Attention Patterns in Autoregressive Generation



Key Insights

[SOS] Persistent conditioning signal

Token Attends to SOS for global context

Bidirectional attention flow

Causal masking prevents future information leakage

Generation Process: Each token uses SOS as initialization context